DOCUMENT RESUME

ED 189 164 TM 800 335

AUTHOR Smith, Jeffrey K.
TITLE The Role of Measurement in the Process of Instruction.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ERIC-TM-70
PUB DATE Dec 79
NOTE 33p.
AVAILABLE FROM ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, NJ 08541 ($3.50)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Testing: Elementary Secondary Education: *Group Testing: *Informal Assessment: *Instruction: Instructional Improvement: *Measurement: *Standardized Tests: Student Evaluation: Test Construction: *Test Interpretation: Test Selection
IDENTIFIERS *Teacher Developed Tests

ABSTRACT

Educators are more interested in instruction and learning than in testing per se. If evaluation, the process of gathering information for instructional improvement, could be accomplished without the quantitative, formal processes of measurement and testing, there would be no need for them. Unfortunately, informal processes are more inefficient, inaccurate, incomplete, and biased, than testing and measurement, especially with groups. Educators should thus be assertive, knowledgable consumers of standardized tests: if these tests are inappropriate, well-designed classroom tests can be useful. Student test anxiety is largely due to the consequence of testing rather than the activity itself. (Guidelines for selecting standardized tests, developing classroom tests, and a glossary of twenty measurement terms are included). (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189164

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ERIC® REPORT 70

THE ROLE OF MEASUREMENT IN THE PROCESS OF INSTRUCTION

by Jeffrey K. Smith

ERIC® ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08541

TM 800335

The material in this publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Prior to publication, the manuscript was submitted to qualified professionals for critical review and determination of professional competence. This publication has met such standards. Points of view or opinions, however, do not necessarily represent the official view or opinions of either these reviewers or the National Institute of Education.

ERIC Clearinghouse on Tests, Measurement, and Evaluation
Educational Testing Service
Princeton, NJ 08541

December 1979

PREFACE

I don't "believe" in tests. Tests are like carpet tacks; they're not a fit subject for belief or disbelief. For some things, tests are useful; for other things, you need carpet tacks. Rarely, however, do you have to choose between the two. If students could learn effectively, without wasting time and effort, and if we could be certain that they had learned those things that we (or they) wanted them to learn, I would be perfectly happy if they *never* took a test. The purpose of measurement is to *assist* the process of instruction. In order to understand that purpose, we do not need to look at measurement; we need to look at instruction. So let's do that.

What I want to examine here is the instructional process as it exists in classrooms. Most folks call that teaching, but as a former seventh-grade teacher, I know better. Teaching encompasses what the dictionary defines as an "instructional process in classrooms." It also involves other things: Parent conferences. Lunch money. Martin punching out Harold at recess. Gum. The vice principal who smokes cigars in the teachers' cafeteria. These are all wonderful things and an integral part of teaching; they are not, however, instruction, which happens sometime after the pledge to the flag, the loudspeaker announcements, and sundry other classroom ablutions, but before lunch and recess jack your kids up into near frenzy. Daniel Lortie, in an excellent and affordable paperback book called *Schoolteacher*,* says that when teachers are asked what a good day at school means to them, they usually reply that a day when they actually get to do some teaching is a good day. A bad day is a day that is torn apart with constant interruptions. I agree with this perspective and am not going to propose testing children three times a day. But I do believe that testing can be an effective, efficient, and nonthreatening method of gathering information for making instructional decisions about children. In the following pages, I will present a perspective on measurement and the process of instruction as well as some clarifications and suggestions concerning the field of measurement.

December 1979

Jeffrey K. Smith
Rutgers University

*Daniel Lortie, *Schoolteacher*, University of Chicago Press, 1976.

I. INTRODUCTION

Teaching is an ill-defined art. There are aspects to teaching which, while we understand them implicitly, are rarely made explicit. An example of one such aspect is that an agreement exists between the teacher and the learner that teaching should take place. This agreement is often stated formally, as in an apprenticeship: The apprentice agrees to work for the craftsman in return for the training he/she receives. In classroom teaching at the elementary and secondary levels, the agreement is not formal; it is not even entered into voluntarily by the student. Often it is the task of the teacher to work at maintaining this agreement. We usually think of this as "motivation," but it can just as easily be conceptualized as an agreement between teacher and learner that the teacher has something worthwhile to teach, and the learner is willing to make an effort to learn.

Another aspect of teaching not usually considered is that the teacher needs to know where the student is, on several levels, with respect to the content of instruction, in order to make decisions on how to proceed. Not only do we need to know the approximate grade level of a student (in order to find, say, the right basal reader), but we also need to know if the student is having trouble with some vocabulary or perhaps with the syntax in a passage. As trained educators, we can pick up cues on the more microscopic elements needing evaluation without resorting to formal procedures. Sometimes, however, our need for information about students requires moving to somewhat more-formal measures (such as quizzes or worksheets) and sometimes to quite-formal measures (such as the Illinois Test of Psycholinguistic Ability).

Some of the latitude that teachers have traditionally possessed, with respect to the level of formality necessary for obtaining information on students, has been removed by local, state, and even federal authority. Districtwide testing programs, statewide testing, and federal mandates such as Title I regulations and Public Law 94-142 have substantially increased the amount of formal evaluation taking place in classrooms. Organizations such as the National Educational Association have opposed this shift. Rather than stating my position on this situation at this point, I would prefer to let my view evolve over the course of this paper.

Irrespective of one's position, it has become necessary for teachers (*and* parents) to become increasingly aware of the process of evaluating students' learning and to become more sophisticated users of formalized evaluative techniques. The purposes of this paper are to increase awareness of and promote sophistication about the role of measurement in the process of instruction. This paper is intended primarily for classroom teachers. It is also intended for those administrators, school board members, reading coordinators, and

teacher association members who are involved in the selection of standardized tests. It may also be useful for parents, but that audience is not of direct concern here.

Why this paper and why now? There are literally hundreds of tests-and-measurement publications available (Buros, 1978). This one is based on several assumptions about the reader that I believe make it especially useful for the practicing classroom teacher. The assumptions are that:

1. The reader is an intelligent and dedicated professional who sees a need for the continuing education of all professionals, including teachers.
2. The reader is concerned about the testing and other information gathering that the reader engages in and that others mandate him/her to do.
3. The reader is inherently more interested in Shakespeare, science demonstrations, and toothless smiles on small faces than the standard error of measurement.
4. The reader has a limited amount of time that he/she wishes to devote to this topic.

The sections that follow include discussions of measurement and classroom instruction, standardized tests and testing terms, and some considerations for constructing your own tests. The fifth section contains some final thoughts.

II. MEASUREMENT AND THE PROCESS OF INSTRUCTION

"All the world is a stage" is a particularly useful concept of life for a playwright. As a measurement specialist, however, I would rephrase Shakespeare as follows:

All the world is a _____.

A. Series of evaluations.
B. Multiple-choice test.
C. No. 2 lead pencil.
D. Trick question.
E. None of the above

The nonsense above serves well as a caveat for this section: Anyone looking for a rationale for the use of measurement in instruction should be suspicious of measurement specialists. (Measurement specialists *liked* taking tests as children.) Therefore, do not accept what follows simply because it comes from a measurement specialist; if the arguments are not persuasive, use your judgment as a professional educator. Later, in the discussion of the specifics of measurement, I will occasionally ask you to take my word for something, but I am not going to ask that of you now.

The Nature of Instruction

Even in the simplest of settings, instruction is highly complex. Fortunately, for our purposes here, we do not need to address all the aspects of instruction. We only need to look at instruction as it relates to measurement.

To begin, have you ever wondered why it is so much easier to explain something to someone in person than it is to write the explanation? In part, of course, it is easier because one can demonstrate in person and cannot on paper. Equally important, in a person-to-person situation, the teacher can see and hear the student's reactions—a nod of the head, a quizzical look, a correct answer to a question, or the words "I don't understand." The diversity of the information the student can communicate easily in this one-on-one setting is impressive:

1. I don't understand that.
2. Could you say that again?
3. Could you give me an example?
4. I already know this, let's move on.

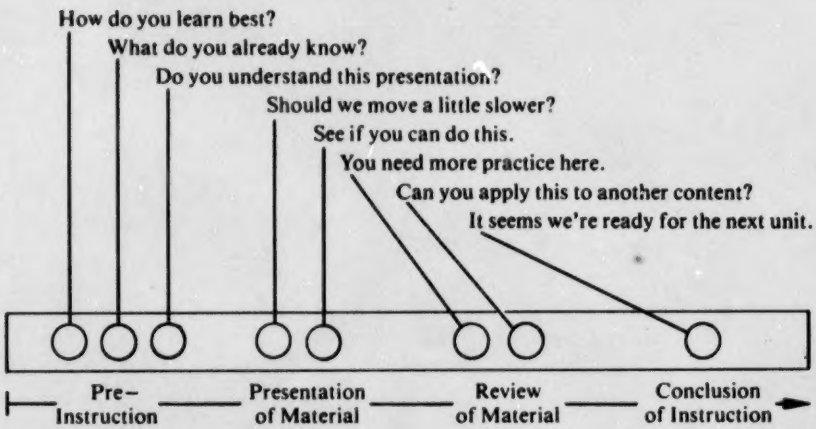
5. Could you show me some other way?
6. What does that word mean?
7. Can you relate this to something I can understand?
8. Could we go slower (faster)?
9. Let me try it to see if I understand.
10. I need some practice to make sure I can do this.
11. This is the way I learn best.
12. I understand.

These questions and statements are not instruction in and of themselves; they facilitate the process of instruction. If one looks carefully at the list, one can see that certain statements occur *early* in the process of instruction (or even *before* instruction begins), some occur typically *during* instruction, and some at or near the *end* of instruction.

From this, we may be able to extract a general principle* concerning instruction: "Throughout the process of instruction, information about the learner and his/her learning facilitates instruction." A diagram of this principle is presented in Figure 1.

Figure 1

Information Concerning Learners and Learning


How do you learn best?
What do you already know?
Do you understand this presentation?
Should we move a little slower?
See if you can do this.
You need more practice here.
Can you apply this to another content?
It seems we're ready for the next unit.
Pre–Instruction
Presentation of Material
Review of Material
Conclusion of Instruction

The conceptualization of instruction presented in Figure 1 is a temporal one; that is, instruction is abstracted as a process with a beginning, middle, and end but without any content. The point is that the content of the instructional process is somewhat independent of the need for information about the

*Contention may be a better word here than principle.

learner. Regardless of what one chooses to teach, information about the learner and his/her learning is helpful.

Gathering Information Informally

Up to this point, the discussion has been limited to a situation with one teacher and one student. In such a setting, most of the desired information can be obtained through interpersonal communication. But we are rarely fortunate enough to have the opportunity to teach one student at a time. So, as we begin to talk about gathering information for purposes of instruction, it would be useful to discuss instruction in the group, rather than the individual, setting.

In the discussion of the individual setting, all examples of information gathering were informal. But certain types of information (concerning such conditions as learning impairment, visual or hearing difficulty, and aphasia, for example) are rather difficult to assess without standardized procedures. When working with a group of individuals, the gathering of data solely on an informal basis poses several problems. Here are four examples.

Some Problems in Gathering Data Informally on Groups of People: The first problem is inefficiency. It simply is not possible to assess the progress of 5 or 25 pupils using the same method one would use for a single pupil.

The second problem is inaccurate information. If 8 out of 10 heads nod in response to a comprehension-type question, it may be concluded that the group is ready to proceed when, in fact, several children may not be ready at all. Several of the nodders may be trying to please the teacher; one may have misunderstood the question. These possibilities also exist, of course, in a one-on-one situation; it's just easier to catch them on an individual basis.

The third problem is incomplete data. Time considerations limit the frequency of informal information gathering and the number of pupils on whom information can be obtained. For example, if one were interested in assessing multiplication facts, it would be difficult to get a complete assessment that would allow for pinpointing weaknesses if it were done on one pupil at a time for 25 pupils.

The fourth problem—that of bias—may be the most serious of all. The problem of bias is slightly different from the problem of inaccuracy or error. Error is simply being off the mark, sometimes high, other times low, but not *consistently* one or the other. Bias occurs when we err consistently in one direction—for example, when we continually believe a student can do things that he/she, in fact, cannot do or when we consistently sell a student short. It can occur within groups of students and against individuals. Racism and sexism are not the only causes of bias; it can occur against "the low reading group,"

the "morning group" (as opposed to the "afternoon group"), or even the "kids in that corner."

It should be remembered that the bias we are talking about is not overt, blatant, discriminatory behavior. This kind of bias has to do with misinterpretation of subtle communications, quite subconscious. The bias can stem from the best of motivations and may be equally harmful to the pupil whether it takes a negative *or* a positive direction.

To summarize: (1) There exists a need for information about learners and their learning in order to facilitate instruction; (2) the nature of the information can be quite diverse; and (3) informal information gathering suffers from inefficiency, inaccuracy, incompleteness, and bias.

It would be nice to be able to say that all the problems inherent in the informal process could be solved by more-formal procedures. Unfortunately, that is not the case. However, some problems can be ameliorated. We will look at these more-formal procedures next.

Gathering Information Formally

Formal is an unfortunate word. Formal weddings, formal dinners, and formal affairs do not sound nearly as interesting as their *in*formal counterparts. In the context of gathering information for evaluative purposes, "formal" simply means that the information was gathered in a systematic fashion, following procedures that have proved to be useful. We don't mean stuffy and rigid, we do mean not offhand or lackadaisical. In this section, we will be discussing the gathering of information in a formal fashion, predominantly through the use of tests.

The perceptive reader will have noted that use of the term "test" has been assiduously avoided up to this point. The reason for this is that testing is a loaded concept in American education today. For students, testing is equated with being ranked and graded; for teachers, testing is equated with accountability and loss of classroom time; for the public in general, testing is equated with the stratification of individuals based upon inaccurate and biased estimates of narrowly defined cognitive abilities. Thus, testing has a bad name, perhaps deservedly, but bad in any case. Even if this were not the case, testing should be viewed only from the perspective of its contribution to instruction. We have, therefore, avoided the term. Now, however, we must begin to use and define "testing" and some similar terms so that we can be more precise in the discussion that follows. To begin, there is the concept of evaluation. "Evaluation" is a very broad term that can be applied to the activities of an art critic or an auto mechanic. A variety of definitions might be given for evaluation, but since we are using the term in an educational context, we'll use an

educational definition. *"**Evaluation** is the process of gathering information for the purpose of making educational decisions."* The word "process" is very important here. It includes the determination of *what needs to be known, how the information will be gathered, what importance will be assigned to various pieces of information,* and *how the information will be used.*

The next term we encounter is "measurement." It is far more limited in scope than evaluation but broader than testing. *"**Measurement** is the process of making a quantitative abstraction of a characteristic of an individual or other object (such as a classroom)."*

By "quantitative abstraction" we simply mean that a characteristic that exists in the real world, such as height, home-run hitting ability, or reading ability, is expressed as a number. The number is an abstraction because it doesn't carry with it all of the richness of the original characteristic. Height is a well-defined characteristic, is easy to measure precisely, and does not lose much when turned into a number. Home-run hitting ability is more of a problem. Should it be defined as home runs in a lifetime? The number of home runs per time at bat? Or the number of home runs per pound of the batsman? Using these different definitions, we obtain different people as outstanding home-run hitters.* Once defined however, home-run hitting ability is easy to measure.

Reading ability is especially difficult to measure. First, it is difficult to define. Second, we don't have a universally accepted metric for it (although grade equivalent is very popular). Third, it isn't directly observable. To be sure, we can observe people reading aloud, but that isn't really what we're interested in. What we *are* interested in is something that exists inside one's head (mind, brain). To measure it, we provide tasks that we believe will require a person to demonstrate the ability we are interested in. It's a tricky business.

To measure mental abilities, characteristics, or states of being, we often resort to "tests." Our third definition: *"A **test** is a task or series of tasks with observable results which are combined and used to estimate an ability, characteristic, or state of being in a person."* Tests are usually paper-and-pencil activites. Of course, this isn't a necessity; a road test for a driver's license is a good example of a non-paper-and-pencil test. The Stanford Binet is a test; so is a quiz. The typical task that we present to students is a question, which testing people refer to as an "item," since many "questions" do not, in fact, have a question mark after them (true-false items, for example).

We have spent some time differentiating the terms "evaluation," "measurement," and "test." In providing these definitions, we have gained an appropriate perspective from which to view testing. As teachers, we are not essentially interested in testing at all! We are interested in evaluation.

---

*Probably Henry Aaron, Babe Ruth, and Ernie Banks, respectively.

Look again, carefully, at the definitions. If evaluation could be accomplished without measurement or testing, there would be no need for either. When informal gathering is sufficient, there is no need to resort to more-formal procedures. Unfortunately, informal procedures are, all too frequently, not sufficient. Recall the four problems concerning informal procedures that were discussed earlier:

1. inefficiency
2. inaccuracy
3. incompleteness
4. bias

The question we must ask here is: Can more-formal procedures, such as testing, alleviate the four problems mentioned above? The answer is: Sometimes. Let's look at the four issues separately.

Inefficiency: Inefficiency seems to be the problem area testing can help most. As mentioned previously, it takes no more time to assess multiplication-fact knowledge for 25 pupils than it does for one (at worst, not much more time). A paper-and-pencil activity can almost always be conducted efficiently in group settings. Some may argue that the information gained through testing is not worthwhile and, therefore, the economy of testing is a false one. We will address the usefulness of the information later; in terms of efficiency alone, testing is a very good proposition except when one is in the actual process of presenting material to students. The pace and flow of instruction is critical to classroom learning; only in unusual cases should this pace be broken by, say, a classroom quiz. In order to check class progress *during* instruction, questioning techniques should be employed. However, at the beginning or end of a presentation, a short quiz that is *not-for-grades* (more on this later) is an excellent and nonthreatening check on student comprehension. Inefficiency is a problem that is frequently resolved through testing.

Inaccuracy: Inaccuracy is a tougher problem to overcome than inefficiency. Measurement people usually deal with inaccuracy under the headings of validity and reliability, which are discussed in the next section. For now, let us consider inaccuracy from a narrower perspective. If a teacher can construct several tasks which he/she feels will require the student to use skills or facts being taught, and if a student responds properly to those tasks, an inference of competence would seem appropriate. In a one-to-one setting, this can be accomplished orally. In a group setting, this is quite difficult to do informally since one cannot go from pupil to pupil requiring the same set of tasks to be performed.

What is often substituted for a thorough one-to-one assessment is an oral assessment of the class as a whole. This leads to improper inferences about

individual pupils. Simply because Johnny responded correctly and Mary nodded while Johnny answered does not mean that Mary understood. The essential point here is that what can be accomplished *informally* on a one-to-one basis can be accomplished *formally* with a group of pupils if we simply provide some structure for the process. Asking a pupil several questions orally and having him/her respond orally is not so different from asking 20 students the same questions and having them write their answers. (Interactive follow-up can occur in the group setting once the evaluation has been completed.) When this process is completed, the teacher will know how competent all students are instead of one. Thus, our information on everyone is based on data and not inferred from smiling faces and nodding heads.

Incompleteness: Testing is very useful in coping with problems of incomplete information. In developing a unit quiz, for example, a teacher can first outline all of the relevant aspects of instruction that he/she wishes to cover (this is discussed in Section IV). In this way, all students are exposed to all relevant tasks. Thus, our information is not only more accurate but more complete.

Bias: The issue of bias is complex since bias affects various sub-populations differently. Here we will compare the bias produced by formal evaluative techniques with that produced by informal evaluative techniques. In testing, we often encounter bias in the way we interpret the results of a test. For example, we infer lack of intelligence when, in fact, a low test score may be due to lack of familiarity with the culture on which the test is based. Two students with identical scores may be quite different in the ability being tested.

Informal evaluation allows for the possibility of ameliorating this bias through a sensitive modification of the data gathering. That is, one can trace apparent inabilities to their causes and, thus, produce results that are less biased. On the other hand, this may produce even more bias in the final analysis because of expectation of ability or inability, which is a problem of informal techniques. Although bias is still considerable and perplexing in formal procedures, the potential for bias probably looms even larger in informal assessment (especially when groups of students are under consideration).

Some Recommendations for Using Measurement

Having presented several pages of apologia for the use of tests, let me now present a set of recommendations concerning the testing of students:

1. Don't use tests for grades.
2. Don't have students review for tests.
3. Give students their results immediately.

The key to understanding these statements is this: Learning is what is important; testing is only there to help.

Let's examine the three points one at a time.

Have you ever wondered why people love riddles and crossword puzzles and hate tests? I would contend it has nothing to do with the activity itself but with its consequences. If grading and testing received a divorce on the grounds of mutual incompatibility, then maybe people would not be so averse to taking tests. Of course, this raises the issue of how to assign grades, which is a "whole other topic." If you must use tests for grades, then simply use a midterm and a final; the use of quizzes or quarterlies is counterproductive. Grades are not important; learning is important.

This brings us to the second point. Don't have students review for tests; have them review *because* of tests. That is, let the test results guide you and your students in the review of material.

Third, we should provide students with results immediately. How? Simple. On short-answer, multiple-choice, or true/false tests, it is easy to have students record two copies of their answers. They hand one in and keep the second, and the class discusses the test that day or the next. This is important. Students who have just completed a number of mental tasks need information about their performance. Problems must be worked out and misperceptions cleared up before they are set.

A Model of Measurement and Instruction

These recommendations fit into a general model or framework of instruction and measurement that can be presented by posing the following series of questions concerning instruction that measurement can answer:

Where do we begin? and How do we proceed? Before instruction begins, we need to know where students are and how we can help them to proceed. There are a variety of ways to gather the necessary information to answer these two questions. There are existing records concerning prior performance: teacher recommendations and comments, grades, and previous test scores.* A teacher may also take an inventory on a student at the beginning of the year, having the student read aloud, work math problems, and the like, in order to assess the student's current status. Finally, the teacher can administer standardized tests to the class as a whole. It is important to make sure that these tests are thorough enough to provide useful information for instructional purposes. (Tests with reliabilities of under .90 rarely meet this criterion.) On what basis

*Previous test scores can be particularly useful if they are presented to the teacher in a useful fashion. All too often, test scores are organized by the teacher's roster of the year before. Standardized test scores should be reorganized into the current teacher's class. This could be accomplished by test publishers or scoring services over the summer.

should the teacher decide to rely on any one of the three suggested sources of information? Basically, the teacher should ask, "Do I have enough information to plan instruction for this child?" If not, an informal or formal procedure should be employed. How do you choose between these options? Well, for how many students do you need additional information? How comfortable are you with your informal techniques? If you have more than a few students, or if you are not comfortable with informal procedures, then you are probably better off with a formal procedure of some type.

Soon after instruction has begun, the third question will arise: How much have the students learned so far? A corollary question is: "Are we ready to move on?" Usually this question can be answered with a quiz (which measurement people call a "formative test"). One aspect of classroom instruction that is almost as universal as it is counterproductive is having students study for a quiz. Let the quiz results *guide* their study, not motivate it.

At the end of the academic grading period, semester, or year, we ask, "How did we do overall?" Frequently, there is a concomitant need for assigning grades. At this point, a final exam or test (what measurement people call a "summative test") might be appropriate. I, personally, don't like to see grades tied to test scores, but I must admit I don't have much in the way of alternatives—not without engaging in philosophical discussions about the nature and purpose of American education. At the end of the year, it is often useful, for systemwide evaluation purposes, to administer a standardized test of some type. What is often ignored in school systems is the value of these scores in planning for the next academic year. (See the footnote on p. 10.)

To summarize the perspective on measurement and instruction presented here, the following statements are appropriate:

1. As educators, we are interested in instruction and learning, not in testing per se.
2. Evaluation is an essential aspect of the instructional process.
3. Tests, as a form of measurement, can be very useful in evaluation. They have properties that often make them more useful than informal procedures.
4. We should use tests for a clearly identified purpose, and when we do use tests, we should be certain that testing is the *best* way to obtain the information.
5. Many of the problems of test anxiety are not attributable to tests but to the consequences of tests, which can be modified.

Having concluded this discussion of the role of measurement in instruction, we leave the arena of persuasion and move into those areas more directly concerning technical expertise.

11

III. UNDERSTANDING STANDARDIZED TESTS

In this section, I will provide you with some assistance in selecting and constructing tests. We will look at how to select a standardized test, interpret its scores, and read through measurement jargon. Then, in the next section, we will run through some practical steps for test construction.

Selecting a Standardized Test

There are literally thousands of standardized tests available to educators: some of them are quite good; many of them are very poor. The question that occurs to the practitioner is, "Should I use a standardized test, and if so, which one should I use?" Actually, there is a question that should arise prior to that one: "What do I want to measure and why?"

There are two reasons for asking this question. The first is that it will help you decide whether you want a standardized test or a locally constructed one (or an informal procedure). Second, it will help you decide which standardized test you want, if that should be your choice. But first we need to address the issue of standardized vs. locally constructed tests.

The classroom teacher may say at this point, "All of these decisions are out of my hands." That may be true at the individual classroom level, although there are hundreds of tests designed specifically for classroom use and many administrators would be willing to purchase such tests if a solid argument for their use could be made. Also, most school districts include teacher representatives in test-purchasing decisions. It may well be the case, however, that the next section is more appropriate for administrators.

Advantages and Disadvantages of Standardized Tests

Although standardized tests have come under a considerable amount of well-deserved fire recently, they, in fact, have many advantages when compared to locally constructed tests. Consider the following:

1. They are already written and require no teacher or staff time for development.
2. Most of them were written by professionals and have some evidence of quality.
3. Most of them yield scores that you can compare with those of another group.

Of course, standardized tests have some drawbacks:

1. They cost money.
2. Many of them are quite poorly developed and have little evidence of quality.
3. Most of them are not exactly what you had in mind.
4. Many of them provide inappropriate norm groups.

So how should one make a decision? Get help. This is a fairly complex activity, and a few dollars spent for a consultant (not a publisher's representative) can make a substantial difference in the utility of your testing activity. Here are a few questions you will have to answer:

1. How important is it for you to compare your group with a norm group? Norm data help with Title I evaluation or for convincing board members or the public at large of school progress. You can't do that with your own test.
2. How important is it that the test correspond with your curriculum? Very few standardized tests will match your curriculum as closely as you would like.
3. How communicable do the results need to be? This concern is similar to the first. If several individuals or groups need to use the results, a test that is widely known has some advantages.
4. How technical or complex is the trait you are measuring? It is difficult for a local school district to develop, say, an early-screening device for learning problems. There are, however, some notable exceptions to this. (See Naron, 1977.)

Having presented these guidelines, let me reiterate: Get help. Call your local college or university and ask for some assistance from the faculty. Some states have agencies that are designed to help school districts. You might also try your state's department of education.

Before you finally decide on a standardized instrument, be sure to look up the test in Buros' *Mental Measurements Yearbooks.* These reference books have rigorous and very useful reviews of almost every test under the sun. A note on these books: If a test was reviewed in, say, the seventh edition of the series, it probably won't be included in the eighth edition. Most tests that are worth buying are reviewed in Buros. Keep looking.*

*The *Mental Measurements Yearbooks* have been taken over by the University of Nebraska and will continue to be published.

What All Those Words Mean

As you plunge into the world of standardized testing, you are going to be inundated with jargon. Though special terms are often useful, some people in the field are deliberately obscure for the purpose of glossing over weaknesses. This is definitely caveat-emptor time. Below is a nonalphabetical guide to what all the terms mean.

A. *Terms related to different kinds and uses of tests*

Standardized Tests: "Standardized" means the same test has been given under the same conditions to large numbers of people. Just about any test that has ever been given to anybody before someone tried to sell it to you is likely to be called "standardized." That is, the term is used rather loosely. What it should refer to is a test that provides some standards or expectations of performance. What it usually means is that there was a norming group that was given the test.* Be wary. Some of the members of the norming groups for tests you might use for first graders fought in the Great War. Make sure the norms are *current* and *appropriate* for the kind of person you wish to test.

Norm-Referenced Tests: The meaning here is pretty similar to that of standardized tests. Basically, a norm-referenced test is one that yields scores that are interpreted by comparing them with scores earned by other kids on the same test. Telling you that Johnny got a 73 on a test in reading doesn't tell you much. Telling you that Johnny is in the top 10 percent of all fourth-grade students is more informative, if you have a sense of how well students at that level can read. "Norm-referenced" simply means that scores *can be interpreted by comparing them with scores of other people.*

Criterion-Referenced Tests: In 1963, Robert Glaser (Glaser, 1963) introduced the concept of criterion-referenced tests. His basic idea was that sometimes you just want to know whether a kid can do something (like learning multiplication tables) or not, and you don't care how well anybody else does on this. This was a nice bit of insight on Glaser's part. There is, however, one large problem with criterion-referenced tests: If you don't really care how well people as a group do on a test, it is very hard to assess the quality of the test. The reasons for this are not necessarily conceptual. It's just that a criterion-referenced test requires measurement people to address the issue of reliability and validity in a new light, and we're still a little blinded by that light. Beware

*A norming group is supposed to be a representative sample of people who are similar to the people for whom the test is intended. A fourth-grade reading test would have as a norming group a sample of fourth graders drawn from a variety of backgrounds.

of the person who tells you, "We don't need validity or reliability on this test; it's criterion-referenced."

The basic difference between criterion-referenced and norm-referenced tests has to do with how we use the *scores.* In norm referencing, we try to understand scores by comparing people according to behaviors or performance. "Ben is the best bowler in the league," is a norm-referenced statement. For criterion referencing, the score has a meaning related to the test itself, not to how others did on it. "Ben's average is 185," is a criterion-referenced statement. The difference is not so much in the tests as in the referencing system. There are measurement people who contend that the difference is in the *tests.* Although they may have a point, I think the crux of the issue is in the reference system.

A final note: Some people think that criterion-referenced tests have to be short and/or related to the classroom and/or have a cut-off score. None of these conditions are necessary, although they may be true of some criterion-referenced tests.

Mastery Tests: Basically, a mastery test is a criterion-referenced test that has a cut score attached to it. That is, if you are above a predetermined level you are considered to have mastered whatever the test was about. The written test for a driver's license is an excellent example of this. Usually, however, mastery tests are used in instruction, which makes the driver's test a little less illustrative. Of course, people who fail that test usually continue to study, and people who pass it burn their copy of *Rules of the Road,* so it is instructional in that sense. If a teacher gave a quiz on multiplication tables, and if anyone getting more than 90 percent right didn't have to study the tables anymore, then the test would be a mastery test. More about cut scores later.

Content-Referenced and Domain-Referenced Tests: Any idea that yields a glimmer of success in education will be dutifully extended, expanded, and elaborated upon until every photon has been accounted for. This is true with criterion-referenced tests. Content-referenced and domain-referenced tests are marginally different perspectives on the idea of criterion-referenced tests. Think of them as one concept, and you will always be within a degree or two of perfect accuracy.

Objective Tests/Subjective Tests: This sense of *objective* means fair and impartial; it is to be contrasted with subjective. No test is objective; it only aspires to be. Tests are considered to be more objective if:

1. An examinee gets the same score from two different graders.
2. The conditions for tests are the same for everyone.
3. The items mean the same thing to all people.

The third criterion is the one we can never really satisfy. People have different lives and cognitive processes, and the same stimuli can easily mean different things to different people. Do not despair over this situation; even though it means our tests can never be completely objective, it is what keeps us from all being Calvin Coolidge (that is, the fact that we are all different). Hence, the price is not too great.

Formative Tests/Summative Tests: The difference between formative and summative tests really lies in how the tests are used. Formative is used, as a term, in the same sense as "formative years"; there is a developmental or instructional aspect to it. A formative test is one in which the results are used to make decisions about the future instruction of the student. A summative test is used to make a summary statement about a student (such as giving him/her a grade). As you can see, it is difficult to ascertain whether a test is formative or summative until you know what is to be done with it. Typically, though, such tests as final exams, certifying exams, and quizzes used only for grading tend to be more summative, while diagnostic measures and quizzes used for instruction are considered formative.

Diagnostic Tests: Diagnostic tests are a special type of formative test. They are specifically designed to address a question related to a specific aspect of instruction, such as "Does this student have visual difficulties?" or "Does this student need work on letter-sound relations?" The field of diagnostic testing is quite large and really deserves more attention than can be given here (see Rapaport, Gill, and Schafer, 1968).

Cognitive/Affective/Psychomotor: These three terms are convenient ways to classify tests according to the types of things they measure. Cognitive tests measure people's mental abilities (we'll quibble over aptitude and achievement later). Psychomotor measures involve directed physical action on the part of the subject in response to the stimulus. Affective measures tap the subject's attitudes, opinions, and state of mind.

Achievement/Aptitude Tests: Achievement tests are designed to measure proficiency in subjects a person has been taught. Aptitude tests are designed to allow predictions of future achievement. Now, one good way to predict future achievement is to look at present achievement; therefore, many achievement tests serve well as aptitude tests. Measurement specialists love to argue over whether there really is such a thing as aptitude. The issue is not in imminent danger of resolution.

A brief note on the terms related to kinds of measures: A test can be objective/achievement/formative/criterion-referenced/cognitive/mastery all at the same time! A spelling quiz with a score that determined whether a student had to do more spelling work would fit all of those categories. It is useful

to think of a particular test (and usage) and run it through the list of terms to see where it fits.

B. *Terms related to various kinds of scores*

When you receive a student's results on a standardized test, you are likely to run into any number of bizarre-looking scores. A little later, we'll talk about how to interpret them. Here we'll just define them.

Raw Score: Basically, a raw score is the number of items a student gets right. Some testing organizations use what is called "formula scoring," which subtracts a fraction from the number right for each answer that is guessed wrong. Rarely does any scoring system penalize guessing to the point where one shouldn't guess. Usually it is an attempt to neutralize guessing.*

Percent Correct: This is the number of correct answers divided by the total number of items. This is *not at all* similar to a centile or percentile.

Percentile Rank/Centile Rank: These words mean the same thing. A percentile rank tells you what percent of the people in the norm group fell below this student's score. (For example, a percentile rank of 84 means that 84 percent of the norming group fell below the score) If the student in question isn't really a member of this group (such as a seventh grader being compared with fourth graders), the percentile is somewhat less meaningful.†

Stanine: Stanine is an abbreviation of "standard nine," and is a score reported on a scale that is divided into nine segments. Each such score is expressed as a number from one to nine. A stanine of 1 means that a student's score was in the bottom four percent of the norming group; a 2 means the score was between the fifth percentile and the eleventh; a 3 means between twelfth and twenty-third; a 4 means between twenty-fourth and fortieth; a 5 means between forty-first and fifty-ninth; a 6 means between sixtieth and seventy-sixth; a 7 means between seventy-seventh and eighty-eighth; an 8 means between eighty-ninth and ninety-sixth; and a 9 means the top four percent.

Standard Score: A standard score is a score that has been converted from an original raw score, usually by transforming the mean and standard deviation of the raw score. SAT scores and IQ scores are good examples of standard scores. Because it is easier to work with scores that have been converted, most

*This is done by subtracting 1/number of options for each wrong answer from total correct answers.

†It does tell you how the student did compared to fourth graders, if that is of interest to you.

test publishers derive their own standard scores. The problem with standard scores, however, is that they are meaningless unless the test publisher provides information on how to interpret them.

Grade-Equivalent Score: Although this type of score is used in many elementary schools, it has a number of shortcomings. A grade equivalent of 4.5 indicates the average performance of a fourth-grade pupil in the fifth month of school. The main problem with grade equivalents is that there is little evidence to suggest that kids march along one month at a time in their academic development. Furthermore, students' average growth, in general, from third grade to fourth grade may be much greater than from fourth to fifth, but we treat it as if it were the same.

Normal Curve Equivalent (NCE): NCEs are a standard score with a mean of 50 and a standard deviation of 21.06. They were developed to avoid technical difficulties associated with grade equivalents and percentile ranks (you can't properly take averages of grade equivalents or percentile ranks). NCEs are very similar to percentile ranks, although they tend to be more moderate at extremely high and low levels of performance.

C. *Terms related to test quality*

"Valid and reliable" is a phrase that appears in almost any discussion of testing. Since these terms frequently involve numbers, it is common for people to ignore the evidence that a test publisher presents on these issues and just look for a concluding phrase something like ". . . therefore the validity and reliability of this measure is well-established." Since that type of activity makes me shudder, I'm going to present a little longer discussion than usual in this section and explain what these terms *really* mean.

Validity: To begin, validity is all we really care about. If a test is valid, it *has* to be reliable.* The reason measurement specialists talk about reliability so much is that it is easier to calculate. Simply speaking, a test is valid if it measures what you want it to measure. As is true with so many of our other testing concepts, the validity of a test depends upon how it is used. Technically, it is more proper to talk about the validity of a particular application of the test than of the validity of the test itself. This borders on pedantry, so we'll go along with convention and talk about "the validity of a test."

Validity is often confused with the *evidence* of validity. A test could have no validity evidence at all and still be the most valid test ever constructed. It's similar to guilt in a criminal proceeding. The guilt or innocence of a person ex-

*It's like the relationship of "antique" to "old."

ists regardless of the evidence the DA can muster. As the evidence builds up, the jury becomes more and more certain of the guilt of the suspect. However, a person could be guilty of something and there could be little or no evidence of it. When validity evidence is presented in a test's technical manual, it should be viewed as an *argument* for the validity of the test. The evidence has to be weighed and a judgment rendered.

There is one important difference between guilt and validity. Guilt is often viewed as a dichotomous (either-or) situation, whereas validity exists on more of a continuum. That is, it is reasonable to talk about test X being *more* valid than test Y for a particular use. Let's look at some of the types of evidence that people present for validity. The most common type of validity evidence consists of correlating the test in question with another, well-established test that purports to measure the same thing. Sometimes, instead of using another test, people use grades or teacher ratings or some other index. Whenever we compare a test with another measure in this fashion, it is called *concurrent* or sometimes, *criterion-related* validity.

A second type of validity has to do with the nature of the questions on the test. In essence, we are asking, "Are these items a reasonable subset of the total pool of items that might be used to measure this trait?" If the answer is "yes," we are establishing what is called *content* validity.

A third type of validity addresses the question, "Does this test represent a reasonable way to conceptualize the trait we are trying to measure?" This is a little trickier to understand than the other two. For example, we could look at the Stanford-Binet IQ Scales and ask, "Is this what we mean by intelligence?" This type of validity is called *construct* validity. It is determined through the accumulation of research and development of theory in an area and is difficult to determine in a single study.

In presenting validity evidence, test authors are likely to present a lot of statistics that are difficult for the average educator to comprehend. Should you find this to be the case, I have two suggestions:

1. Get help.
2. Read Buros.

Reliability: Reliability is easier to talk about than validity. Reliability is a way to assess the accuracy of a measure. The simplest way to explain reliability is to imagine giving a test twice, about 10 days apart. The reliability coefficient would be an index of how similar the results are in the two administrations (it is actually the correlation between the two sets of scores). If you are using a test to make decisions about individual students (instead of, say, for program evaluation), the reliability coefficient should be above .90, and it would be much better if it were above .94. Many tests that are sold for individual testing

do not even approach these standards. Why should the reliability be so high? Because if it is lower, there is far too much chance for error in a student's score. I will explain more about this under "standard error of measurement." A few final words on reliability: There are different ways to measure reliability (some of which only involve a single administration of the test). Some of the more modern approaches to test accuracy do not use reliability coefficients for a test as a whole but do provide standard errors of measurement for every possible score (this is the case with the increasingly popular Rasch Model). This is fine; in fact, it's superior to a single reliability index. Remember, all tests should provide some index of score accuracy.

Standard error of measurement: Although most test publishers and measurement specialists focus on the idea of reliability, it is really the standard error of measurement (SEM) that is of concern to test users. The SEM tells us just how far away from the truth the student's score might be. Error here does not mean mistake; it means uncertainty. If we could give a test to a student a thousand times and take his/her average score, we would have a good guess at his/her "true" score on this test. But since we usually give a test only once, we need an index of how far from typical performance this particular score might be for this student. If we take the SEM, double it, and add it to the student's score, it will tell us how far off on the low side our observed score *might* be. If we then subtract twice the SEM from the student's score, this will tell us how far off on the high side we might be. We can be about 95 percent sure that a student's true score will be in this range.*

We mentioned before that one should look for reliabilities of .90 and above. Actually, a better procedure is to find the SEM and then multiply it by *four.* This will tell you the range of possible scores that you will encounter (+2 and −2) for each student. For example, some standardized reading tests have SEMs of .7 grade-equivalent years. Multiplying by 4, we get a range of almost 3 grade-equivalent years. Using this SEM, an estimated grade equivalent of 3.6 might be as high as 5.0 or as low as 2.2. Clearly, this is unacceptable for making instructional decisions about students.

Some tests give one SEM for the test as a whole; others give a different SEM for each possible score, smaller toward the middle scores and larger toward the extreme scores. This latter procedure is generally preferable, since it is a better reflection of the reality of the situation. It is critical for you, the test user, to be sure that the range given by four times the SEM yields a measure that is accurate enough for you.

*This isn't a very precise description of the procedure, but it's close enough for most purposes.

What Do I Do With the Results?

In some respects, if you are asking this question you probably shouldn't be testing in the first place. That is, when you are testing you should know in advance what information you are looking for and how (in what form) you are going to receive it. In order to accomplish this, first read the teacher's manual if one comes with the test. It was probably written with someone like you in mind. If the manual isn't clear, call or write the test publisher or author. Don't be afraid to question what is in the manual. There are no magic scores in the field of measurement. Although some test scores need to be interpreted in combination with others, there are few, if any, scores that cannot be interpreted by a reasonably intelligent, experienced educator.

The computer printout you receive may be difficult to decipher, so make sure you read any accompanying material carefully. If your printout is impossible to read, complain to the publishers. They will help you understand the printout and may change future versions if they receive enough complaints. You might also get some help from your school's test coordinator or the publisher's representative. You should *expect* a test salesperson or publisher to be able to explain *clearly* what you are getting and how to use it. If you can't get a sufficiently clear explanation, *don't use the test.* To summarize, three points:

1. Know what you're getting before you get it.
2. Read the manual or printout thoroughly.
3. If necessary, get an explanation from the people who sold you the test in the first place.

Now on to some suggestions for developing your own tests.

IV. DEVELOPING CLASSROOM TESTS

If you can't find a standardized test that is appropriate for your needs, you may want to develop your own test. There are a variety of excellent texts that can provide suggestions on how to do this (see Thorndike & Hagen, 1977; Gronlund, 1976; Bloom, Hastings, & Madaus, 1971; Wick, 1973).

In this section, I want to briefly outline one approach that I have found to be particularly useful. If this method doesn't seem useful to you, you might consult one of the works listed.

There is nothing magic about what follows: It can be found in some form in most texts. I am presenting it here because if I have gotten you to stay with me this far, you may profit from these ideas even though you may have run across them before.

This is the rationale for what is presented here: You have to know what you put into a test in order to understand what you get out of it. Teachers need to be very careful about the design of a test in order to have confidence in the results.

In order to present these ideas, it might be helpful to use an example. We begin with a need for information about how our students are learning. (If we don't need information, we don't need a test!) Let's say that we have been teaching a social studies unit on different levels of government for three or four weeks, and there is one more week allocated for instruction in this area. It occurs to us that this last week of instruction would be most profitable if we had a good idea of which students already comprehended what material. In essence, we have answered the first question in developing a test: Is this test necessary?

This question might be expanded to: What do I want to get out of this test? The more precisely this question can be answered, the easier the rest becomes and the more useful the test will be. This is worth focusing on a little more closely. Earlier, we said we wanted a "good idea of which students already comprehend what material." But what does this mean? We need a *method* for specifying the information we want from our test. One way to do this is to use a content-behavior matrix. A detailed discussion on developing such a matrix can be found in Bloom, Hastings, and Madaus (1971). The essential idea of a content-behavior matrix is to separate what we want students to be able to do (behavior) from the material or subject matter we want them to do it with (content). For instance, with respect to the social studies unit we might be interested in the following behaviors:

1. Defining terms
2. Understanding the relationships among various elective offices

REFERENCES

Bloom, B. S., Hastings, J. T. & Madaus, G.F. *Handbook on formative and summative evaluation of student learning.* New York: McGraw-Hill, 1971.

Buros, Oscar K. *The eighth mental measurements yearbook.* Highland Park, N.J.: Gryphon Press, 1978.

Glaser, G. R. Instructional technology and the measurement of learning outcomes. *American Psychologist,* 1963, 18, 519-521.

Gronlund, N. E. *Measurement and evaluation in teaching.* (3rd. ed.) New York: Macmillan Publishing Co., Inc., 1976.

Lortie, D. *Schoolteacher.* Chicago: Univ. of Chicago Press, 1976.

Naron, N. K. *The Chicago early project: first year report.* Chicago: Chicago Board of Education, 1977.

Rapaport, D., Gill, M. M., & Schafer, R. *Diagnostic psychological testing.* rev. ed.: Holt, R. R. (Ed.) New York: International Universities Press, 1968.

Thorndike, R. L. & Hagen, E. *Measurement and evaluation in psychology and education.* New York: John Wiley & Sons, 1977.

Wick, J. W. *Educational measurement.* Columbus, Ohio: Merrill Publishing Co., 1973.

3. Applying the concept of checks and balances

These might be the content areas we are interested in:

1. City/municipal government
2. County government
3. State government
4. Federal government

Now, all these behaviors may not be related to all the content areas, but if we construct a matrix, all the possibilities will be apparent, and we can examine them to see which are important to us.

This has been done in Figure 2.

Figure 2

Content-Behavior Matrix

	City	County	State	Federal
Defining terms	$X_{.05}$	$X_{.05}$	$X_{.05}$	$X_{.05}$
Understanding relationships	$X_{.10}$		$X_{.25}$	$X_{.25}$
Applying checks and balances			$X_{.10}$	$X_{.10}$

There are 12 cells in Figure 2, and an analysis of the combination of content and behavior has suggested that nine of these cells are important for our test. We decided that at the county level we were only interested in terms, and that applying checks and balances at the city level was of little interest. However, not all of these cells are equally important. It may be that we are primarily interested in our students' understanding of the relationships among elected offices at the state and federal levels. We might assign 25 percent of the test to each of these categories (50 percent of the total). Then we might decide that terms are worth 5 percent each (20 percent of the total). Of the remaining three cells, we might decide to allocate 10 percent to each. This accounts for 100 percent of our test.

These decisions are somewhat arbitrary. What this activity of assigning weights to various aspects of a test requires is that a teacher specify and quantify what is important to his/her instruction. It is clear that this matrix approach does result in a fairly precise statement of what will be obtained from the test.

At this point, we are ready for the second question: What kinds of items should I use? My answer here is simple: multiple-choice and short-answer.

Occasionally, I hear a good argument for essay questions. Frequently, I hear a bad argument for essay questions. The bad argument has to do with getting students to organize thoughts and communicate them clearly. I am all for teaching students to do this. I am so much in favor of teaching this skill that I don't think it should be used as a means of measuring something else. In this example, organizing and communicating thought was *not* listed as something we were interested in. If we are interested in this skill, we should state that explicitly, and more important, we should *teach* students how to do it. *Then,* we can test it.

The good argument for using essay questions is the same as the bad argument. The only difference is that in the good argument, we state explicitly that we are interested in the skills, *and* we address them in instruction.

Two other types of questions are possible: matching, which I don't care for much; and true-false, for which I won't even listen to arguments. (Half of the people who don't know the answer to a true-false item get it right anyway.) Matching items are good for tying capitals to states, exports to countries, and inventors to inventions. These may be worthwhile, but I can't seem to get excited about them.

This leaves us with multiple-choice and short-answer items. Short-answer is a good item type because it all but eliminates guessing as a factor. Unfortunately, it is often hard to measure a good range of abilities with short-answer items. Also, scoring can occasionally be ambiguous.

Multiple-choice items have many advantages: Scoring is quite simple and completely objective (in that two people will score the test the same way); a broad range of abilities can be tapped; and the item format is widely used and generally understood by most students. It has, however, two substantial drawbacks. The first is that it is possible to guess the correct answer. One can never be *certain* that a correct response indicates competence on the question. The second problem is that we cannot tap production, but only recognition of correct responses. Therefore, a mixture of short-answer and multiple-choice items is usually a useful format.

Having decided upon item format, the next question is: How many items? The answer here is usually dictated by practical terms. How much time is available? About one minute per item is usually a good amount of time to allot. Let's say that we decide to use 30 items for our example. In order to decide how many items to write for each cell, we simply need to multiply the proportional weight of each cell (Figure 2) by 30, the number of items we need. Often this will involve some rounding. Don't worry if you end up with 38 or 33 items instead of 30; what is important is that the final distribution of items among cells is the way you wanted it to be. For example, in our government test we would have three items ($30 \times .10$) on "Understanding relationships/City" and seven or eight ($30 \times .25$) on "Understanding relationships/State (see Figure 2).

We have reached the next-to-last question: How do I write items? The answer to this question is worth a book. I recommend Gronlund (1973) or whatever good introductory measurement text is available to you. The issue of item writing is too important to receive a cursory examination. Let me make one suggestion: strive for clarity. If a student knows what you *want* him/her to know on a particular item, your goal should be to make it impossible for that student to get the item wrong. Conversely, you should try to make it impossible for a student who *doesn't* know the content of an item to get the right answer. To me, the first goal is paramount. For more information, invest some time in reading about item writing; it will be well worth it.

Once the items are written, putting the test together and administering it is a fairly straightforward activity. One suggestion here: Have students make two copies of their answers. They can turn one copy in after completing the test, and when all students have finished, you can go over the test with them immediately while their responses are still fresh in their minds.

The final question is: What do I do with the results? To answer this, we need to return to the need for the test. Recall that we had one week of instruction left and were looking for the most profitable way to spend it. To begin our analysis of the test results, we might ask: What does the entire class seem to need help with? Are there any questions, or cells, that most students did not perform well on? A good way to investigate this is to organize all of the questions by cell and then list them across the top of a sheet of paper. Next, arrange the students' scores from the highest to the lowest. List their names down the side of the paper. This will create a matrix as in Figure 3 on page 26. Now mark an "X" in each cell where a student missed an item. This is simple to do and allows for a quick inspection of performance on a cell-by-cell or item-by-item basis.

Having determined the strengths and weaknesses of the class, you can now do a student-by-student analysis. Perhaps students can be put into groups according to common difficulties (for example, six students may have had trouble with federal checks and balances; they might work together). In general, this kind of analysis allows for statements about:

1. Needs of the class as a whole.
2. Needs of groups of students.
3. Needs of individual students.

Figure 3

Analysis of Test Results

	Terms City		Terms County		Terms State		Terms Federal		Relationships City			Relationships State			Relationships Federal					
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	...
Ben.....									X				X							
Lauren...									X	X	X									
Tim			X	X					X				X					X		
Kevin									X		X		X		X					
Jane.....			X							X		X					X			
Greg.....			X	X			X		X	X										
Madelyn..			X	X					X		X	X	X	X						
Ralph....			X	X	X		X		X	X			X			X				
Brian		X				X	X		X	X	X		X	X	X					
Randy ...	X		X	X		X			X		X		X		X	X				
Rachel ...																				
.																				
.																				
.																				
.																				

By examining the pattern of incorrect responses on this partial class diagram, we can make the following conclusions about instruction:

1. Most of the class could use some review of understanding the relationships among elected city offices (items 9-11).
2. A group of students need help with county terms (items 3, 4).
3. Item 13 may be poorly written. Students had much more trouble with it than 12 or 14, which measure the same cell.
4. Kevin has the terms down but has trouble with relationships among elected officials.
5. There are obviously other conclusions that can be made from these data.

We began with a need to know how to proceed with instruction. We concluded with statements that will help us do just that. If this seems neater and tighter than most measurement activities, it is not just happenstance. Usefulness should be a goal, not a fortuitous outcome.*

*Some needs do not lead to such nice turnarounds, but they are still important; for example, the need for program evaluation, districtwide assessment of student progress, and so on. Here the payoff is not quite so rapid and direct as with classroom instruction. The data provide answers to questions such as "Should we change the Title I curriculum?" "Has Centerville solved its bilingual education problem?" We should recognize that these, too, are decisions that require data even though test results may not lead to such clear and immediate benefit to the test takers.

Note that grading students as a result of the test scores was not mentioned. It is a different issue. If the test scores are used for grades, we introduce a host of new elements into what was previously an uncomplicated procedure. Now we have competition, anxiety, stress, and resentment. This is great for training advertising executives, but it's a poor way to teach social studies.

A suggestion before leaving this area: Some of the educators I encounter feel they know their students so well that they could fill in the X's in Figure 3 without giving the test. Even if you aren't that confident, try doing just that some time. If you are about to give a test, make a chart like the one in Figure 3 and guess what the results will be (perhaps by using 0's instead of X's). When the results come in, put in the X's and check your accuracy. The differences between the X's and the 0's are an indication of how much more useful the formal procedure was than informal speculation.

Finally, writing your own test takes time; time that could be spent on other activities. Is writing a test worth the time? is always a consideration. My suggestion is to try the procedures mentioned in this chapter once or twice, and then you'll know what the answer should be in your case.

V. SOME FINAL THOUGHTS

In closing, I would like to reiterate what I believe to be the more salient ideas presented in this discussion.

1. The purpose of measurement (and therefore testing) is to facilitate the instructional process. The utility of any testing activity ought to be clear and demonstrable.
2. Testing is a useful way to gather information to facilitate instruction, especially with groups of students. That is, I am contending that you can use testing effectively to assist instruction (as contrasted with point (1) which contended that it *ought* to do this whenever it is used).
3. The negative aspect of testing, from the student's perspective, is largely a function of the consequences of testing rather than the activity itself.
4. Educators ought to be assertive (even aggressive), knowledgeable consumers of standardized tests. Read Buros, consult a measurement specialist, and talk to the publisher's representative until you are *certain* that you understand what you will be getting out of the test you buy.
5. There are a bundle of measurement terms, but they aren't too hard to understand.
6. Constructing your own classroom tests is a straightforward procedure that can be quite useful in instruction *if* you plan your construction well. Know what you are *getting out of a test* by knowing what you *put into it.*

I began this paper by stating that I don't "believe" in tests. What I do believe in is informed decision making by teachers. This always requires information—sometimes best acquired by testing. If the consequences of testing are not threatening, the testing activity loses much of its oppressive connotation. Don't take this on faith. Try it.